Sentiment Analysis

A Probabilistic Approach

February 5, 2012

Supervised by I. Langbroek (Blauw Research)
Mentored by dr. M. van Someren (Universiteit van Amsterdam)

S. A. Gieske	S. Laan	C. R. Verschoor	D. S. ten Velthuis	A. J. Wiggers
6167667	6036031	10017321	0577642	6036163

Artificial Intelligence
Faculty of Science
Universiteit van Amsterdam

Contents

1 Introduction

With the growth of Social Media, such as Facebook, Twitter and blogs, consumers gained a place to review, rate and recommend products online. This online opinion is important for companies who want to market their products, manage their reputations and identify new opportunities. The process of finding relevant content, filtering out noise and categorization of messages can be automated, making it several times faster than manual processing.

Sentiment analysis is the application of natural language processing, computational linguistics and text analytics to identify and extract subjective information in source materials. It readies data from social media for further analysis. Through use of sentiment analysis, messages can be clustered into categories (e.g. positive, negative and neutral messages).

2 Problem Specification

The dataset was provided by our client, a spreadsheet of 10.000 messages from social media[1]. All messages were manually placed into 5 categories, ranging from very negative to very positive. The goal of this project is to create a method to classify these messages automatically.

2.1 Data Cleaning

A tweet can only contain 140 characters. To still be able to express oneself, slang and emoticons are used. It is also common to find spelling errors in the messages. This increases the value of cleaning the messages before classifying them. Punctuation usually does not add to the sentiment of the message. Therefore, all punctuation except exclamation marks and question marks are ignored. These two do give value to the sentiment of a sentence because it says something about the emotion of the user.

A Dutch stemmer, found in the natural language processing package for Python, enables reducing inflected words to their stem, base or root form. Two words that lead to the same classification, for example "Evangelisch" and "Evangelische", are related to each other. By stemming the last word to its base form "Evangelisch", the amount of different words in the corpus is decreased. This makes classification much easier, AS WILL BE SHOWN IN WAAAR!???

2.2 Data Reduction

The dataset contains mostly tweets. These short messages usually contain a single subject, whereas blog and facebook posts are in general longer messages that may contain several opinions about several subjects. Twitter being the largest source of information, it was decided that only these posts would be used for classification.

[1] Data contained posts from blogs, facebook, hyves, fora and tweets

There are several methods to further decrease the vast amount of data:

- Remove words that occur only once in the corpus.
- Remove duplicate messages
- Remove words that do not contribute to classification (personal pronouns, so-called stopwords, etc.)

Which of these methods would be useful depends on the machine learning method, e.g. an approach that considers grammar to be an important feature for classification makes it impossible to remove words from a sentence without affecting the results.

3 Theory

3.1 Relevant Literature

3.2 Machine Learning

Machine Learning is a branch in the field of artifical intellgence that researches and develops algorithms cabable of improving predictions or behaviors based on a dataset. An algorithm can take example data to capture the pattern in the underlying probability distribution. A succesful algorithm can automatically find these complex patterns in the training data and make intelligent desisions in new (similar) data. The type of machine learning methods that are used in this assignment is called 'supervised learning'. This type requires that the desired decision, in this case the sentiment of the messages, is already known.

3.2.1 Perceptron

A perceptron is an example of a supervised learning algorithm. It is an artificial neuron that has one or more inputs s_i, a threshold function t and an output o. This output can be defined as $o = t(\sum_i(s_i))$

By adjusting the weights of a neuron the algoritme can decrease the margin of the error. This process is repeated until the error is gone or at its smallest. (vertellen hidden layers? miss te ingewikkeld en niet intressant voor ivo)

3.2.2 Artificial Neural Network

An artificial neural network is a computational model with the same functional aspects of biological neural networks. It consists of a group of connected artificial neurons that changes its structure based on internal and external information. Each neuron has one or more inputs, a threshold or threshold function, and an output.

3.2.3 Maximum Entropy

3.2.4 Naive Bayes Classifier

3.2.5 Weighted Probability Sum

3.3 Classification Measures

In order to test how well a specific algorithm performs, a couple of measurements are used. Each will be explained briefly in terms of table **??**.

		Classified Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Table 1: Table of classification classes

3.3.1 Accuracy

3.3.2 Precision

The precision is defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

3.3.3 Recall

3.3.4 F-Measure

4 Usage and User Guide

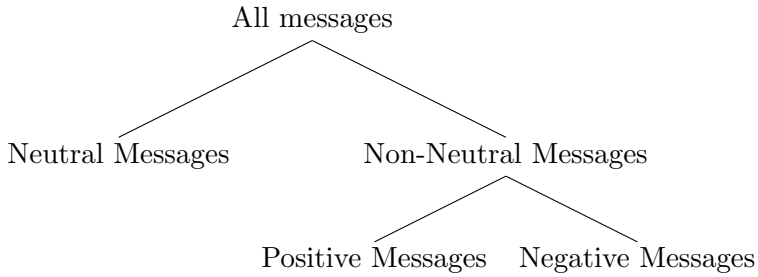
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5 Implementation and Results

5.1 Global Approach

the general approach will be explained first. The idea is to use a binary classifier, i.e. either true or false, multiple times. This way it is possible to first decide whether a message is neutral and then decide between the non-neutral messages whether it is a

positive or negative message. Maybe a tree structure shows the used approach more clearly:



As you can see, using this approach you have to distinguish between two classes at a time. This is very convenient, because there are a lot of techniques that can discriminate between two classes, whereas multi-class-classifiers are less common.

5.2 Binary Classification

In this section, the method that we use to classify a message is described. The approach consists of a few steps:

1. The counting of word frequencies
2. Calculation of word probabilities
3. Calculation of message probabilities
4. Finding a good threshold

First the frequency of each word, i.e. we count how many times it occurs in the training set, is counted. The number The number of times the word occurs in the opposite class, is equivalent to the total number of encounters, minus the encounters in the first class.

When the frequencies of all words[2] are found, the probability for each word can be calculated. This gives an idea of how likely it is to be encountered in a certain class. The formula for this probability is the following:

$$P(word) = \frac{\sum word \in C_1}{\sum word \in C_1 \cup C_2} \tag{1}$$

So the probability that a word is in C_1, the first class, is the number of times it has been encountered in a sentence, which was tagged to be in C_1, divided by the total number of encounters.

Now that all words have probabilities assigned to them, or at least all words in the training set, the probabilities of the sentences can be calculated.

[2] All words in the training set

$$P(s) = \frac{1}{n} \sum_{w \in s} P(w) \tag{2}$$

Where s is the sentence, n the number of words in the sentence and w a word in the sentence. As follows from the formula, the weighted sum of the word probabilities is used.

Now that every sentence has an associated probability, the real machine learning can begin. The goal is to find a threshold for the probabilities; when a sentence probability is higher than a certain amount, it is classified it as C_1, otherwise it belongs to C_2.

5.3 Algoritme 1

5.3.1 Approach

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.3.2 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.3.3 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4 Algoritme 2

5.4.1 Approach

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum

ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4.2 Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

5.4.3 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

6 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

7 Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis vulputate molestie mi ac dignissim. Proin tristique convallis volutpat. Nunc semper erat id tortor fermentum ullamcorper. Donec sed erat quis erat condimentum pellentesque. Donec sed tristique quam. Proin dictum convallis velit a porttitor. Curabitur in tellus tortor. Proin aliquet blandit sagittis. Curabitur vitae mauris ac leo dignissim rhoncus nec ut orci. Praesent vulputate mollis auctor. Aenean in felis diam, quis dictum metus.